

How AI Responds to Common Lung Cancer Questions: ChatGPT vs Google Bard.

Year: 2023 | Citations: 227 | Authors: A. Rahsepar, Neda Tavakoli, G. Kim, Cameron Hassani, F. Abtin

Abstract

Background The recent release of large language models (LLMs) for public use, such as ChatGPT and Google Bard, has opened up a multitude of potential benefits as well as challenges. **Purpose** To evaluate and compare the accuracy and consistency of responses generated by publicly available ChatGPT-3.5 and Google Bard to non-expert questions related to lung cancer prevention, screening, and terminology commonly used in radiology reports based on the recommendation of Lung Imaging Reporting and Data System (Lung-RADS) v2022 from American College of Radiology and Fleischner society. **Materials and Methods** Forty of the exact same questions were created and presented to ChatGPT-3.5 and Google Bard experimental version as well as Bing and Google search engines by three different authors of this paper. Each answer was reviewed by two radiologists for accuracy. Responses were scored as correct, partially correct, incorrect, or unanswered. Consistency was also evaluated among the answers. Here, consistency was defined as the agreement between the three answers provided by ChatGPT-3.5, Google Bard experimental version, Bing, and Google search engines regardless of whether the concept conveyed was correct or incorrect. The accuracy among different tools were evaluated using Stata. **Results** ChatGPT-3.5 answered 120 questions with 85 (70.8%) correct, 14 (11.7%) partially correct, and 21 (17.5%) incorrect. Google Bard did not answer 23 (19.1%) questions. Among the 97 questions answered by Google Bard, 62 (51.7%) were correct, 11 (9.2%) were partially correct, and 24 (20%) were incorrect. Bing answered 120 questions with 74 (61.7%) correct, 13 (10.8%) partially correct, and 33 (27.5%) incorrect. Google search engine answered 120 questions with 66 (55%) correct, 27 (22.5%) partially correct, and 27 (22.5%) incorrect. The ChatGPT-3.5 is more likely to provide correct or partially answer than Google Bard, approximately by 1.5 folds (OR = 1.55, P = 0.004). ChatGPT-3.5 and Google search engine were more likely to be consistent than Google Bard by approximately 7 and 29 folds (OR = 6.65, P = 0.002 for ChatGPT and OR = 28.83, P = 0.002 for Google search engine, respectively). **Conclusion** Although ChatGPT-3.5 had a higher accuracy in comparison with the other tools, neither ChatGPT nor Google Bard, Bing and Google search engines answered all questions correctly and with 100% consistency.